DOCUMENT RESUME

ED 442 270 FL 026 261

AUTHOR Pollard, John Douglas Edward
TITLE Research and Development: A Complex Relationship Part I [and] Part II.
PUB DATE 1999-00-00
NOTE 28p.; Previously published in Language Testing Update: 24 Fall 1998 and 26 1999 by the International Association of Language Testers. Part I presented at the CTELT Conference (Al-Ain, United Arab Emirates, 1998); Part II presented at the Annual Meeting of the Teachers of English to Speakers of Other Languages (TESOL) (33rd, New York, NY, March 9-13, 1999).
PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Applied Linguistics; *Computer Assisted Testing; *Computer Managed Instruction; *English (Second Language); Foreign Countries; *Language Proficiency; *Language Tests; Limited English Speaking; *Oral Language; Second Language Instruction; Second Language Learning; Test Validity; Testing
IDENTIFIERS Messick (Samuel); Saudi Arabia; United Arab Emirates

ABSTRACT

Part 1 of this document describes the background, format, and early groundwork that went into the development of a test sponsored entirely by private enterprise. The discipline imposed by a financial bottom line imposes special pressures but also offers new opportunities. This private enterprise model is a multi-constructional process where language testing research and literature; English-as-a-Second-Language teacher-cum-test developers; a transient pool of teachers; non-language specialists, directors, managers, and accountants; and the ever-present need for performance-reliable second language assessment all play a role. Close interaction between researchers and practitioners is essential. Part 2 explains how opportunities to use data from such a test-in-development led to more substantial and valid research designs that could be executed in-development to improve assessor training and reporting descriptors. There is a need to assess the impact of varied test formats and task features on interaction and to explore how these events vary in different cultural settings. The design of this test was an attempt to maximize the conversational style of interaction through features of computer-assisted test design, including a succession of holistic evaluations incrementing to a final score, reducing the processing and memory burden on the native speaking interlocutor, and positioning the evaluations at junctures that could allow the native speaking interlocutor to avoid sudden pre-closure and closure moves once the assessor has gotten what is needed from the student. Minimizing topic hopping is essential to make the assessment more like a conversation than an interview. (KFT)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 442 270

Research and Development
A Complex Relationship
Part I and Part II

John Douglas Edward Pollard

J. Pollard

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☐ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

BEST COPY AVAILABLE

FL026261

Research and development : a complex relationship – Part I.

This is the first of two papers. It describes the background, format, and early groundwork that went into the development of a test sponsored entirely by private enterprise. While imposing special pressures that sometimes conflict with good testing practice, this environment also offers new opportunities. Tests (and test developers) may be able to survive in such environments, and still maintain closer relations with applied linguistic theory and research than has been suggested.

In *Language Testing Update* Issue 20 (Autumn, 1996) a work-in-progress report was published on a 'live' computer-resourced proficiency assessment known as the 'Five Star Test' being developed in Saudi Arabia. At that time an independent research review was underway at Sheffield Hallam University. There has been some development and a considerable expansion in the use of the test since then.

John Douglas Edward Pollard

Research and development : a complex relationship – Part I.

This is the first of two papers. It describes the background, format, and early groundwork that went into the development of a test sponsored entirely by private enterprise. While imposing special pressures that sometimes conflict with good testing practice, this environment also offers new opportunities. Tests (and test developers) may be able to survive in such environments, and still maintain closer relations with applied linguistic theory and research than has been suggested.

In *Language Testing Update* Issue 20 (Autumn, 1996) a work-in-progress report was published on a 'live' computer-resourced proficiency assessment known as the 'Five Star Test' being developed in Saudi Arabia. At that time an independent research review was underway at Sheffield Hallam University. There has been some development and a considerable expansion in the use of the test since then.

Competing forces.

Test developments on this scale of comprehensiveness or intended scope of use are not typically undertaken in purely commercial settings, and are not normally steered by individuals in professional isolation from other testers. This initiative, however, has been motivated and sponsored as a result of a clear 'market need'; has taken place within tight constraints; and has been steered by a 'single-handed' developer. Remaining faithful to trends supported by language testing research has required a great deal of mediation between the competing pressures of testing heuristics and commercial constraints. Phases of the development have at times been forced ahead by sponsoring stakeholders in ways that have seemed premature and unsatisfactory.

The project has thus far been sustainable in commercial terms. In testing terms, the expanding use increases the need for more rigor, and at the same time provides qualitative and quantitative data to inform the process. It is hoped that by making opportunities to analyse this data there may be scope to combine complimentary roles of research and development, and perhaps inform a wider audience on some of the troublesome issues emerging from *Pandora's box* (McNamara, 1995; 1996).

At the 1998 Language Testing Research Colloquium (LTRC) J. L. D. Clark remarked that we are increasingly asked to produce 'better, quicker and cheaper' tests. This implies front-end financial and commercial 'stakeholders'; sponsors who form a powerful constituency in determining our work – whether it is conducted in a research institution or in society at large. Such stakeholders can only tolerate limited developmental work relative to that which brings *returns on investment*. The *products* they support must have 'face' and 'utilisation' characteristics, fit into the requirements of *customers*, delimiting parameters such as the number of people to be assessed, the time-scales and resources available. This type of stakeholder is referred to only briefly in the 'ethics' issue of *Language Testing*, and may have been underrepresented in our literature as a whole. It may be that the pressures and priorities they bring to bear on language testing can force innovations that would not thrive in purer research environments.

This report does not propose a coherent strategy for reconciling non-specialist decision-makers with applied linguist test developers. The process has been characterised by negotiations, concessions, improvisations and informed risk-taking. It has occasionally required good fortune. To borrow from conversation analysis, it has been a *multi-constructional* process where the participants have been (i) language testing research, (ii) language testing literature, (iii) a TEFL teacher-cum-test developer (myself), (iv) a transient pool of teachers, (v) non-language-specialist directors, managers, and accountants, (vi) a test population, and, last but not least, (vii) a pressing practical need for performance-reliable second language assessment.

The need.

In the early 1990s demographic and economic factors led the government of Saudi Arabia to pressurise businesses into employing more local personnel in an initiative called 'Saudisation'. English is the *lingua franca* in local banking, hotel, aviation, computer and

FT 026261

military establishments, and the medium of training where resources are unavailable in Arabic. This has increased the demand for selection and placement tests in an environment dominated by 'indirect' discrete item approaches. A previous climate of *laise faire* had paid little attention to developments in language testing, but this more recent need to employ and train Saudi personnel has produced a vocal discontent with post-test performances.

At the outset this project involved explaining these inconsistent results to non-linguist decision-makers and proposing suitable alternatives. It was pointed out that language testing was represented by a substantial body of literature, several major associations, and a number of professional journals. If a test for 'high-stake' purposes was to be developed, it would be advisable to offer it to the scrutiny of independent expertise, and present it at international forums. It was suggested that, generally, tests which directly operationalised second-language skills would be more revealing than indirect ones, and more effective if based on needs analysis and trialling procedures which referenced them to learners and target situations rather than idealised language components. The caution was added that 'direct' tests tend to be more resource-intensive.

A test designed to address the whole range of the test target population was eventually approved. It would necessarily be 'broad and general' and would not be immediately adaptable to other purposes such as diagnosing course content or measuring attainment during ELT programmes. (Alderson, Krahnke and Stansfield, 1987: iv; Alderson, Clapham and Wall, 1995:20)

In these preliminary stages care had been taken to 'credential' the initiative, and raise the awareness of managing sponsors to the complexities involved, but this had to be done without getting too specific and risking prejudices that sometimes align the 'theoretical' in opposition to the 'practical'.

The test.

The test was developed to its current version between 1993 and 1995. In 1994 it was transferred to computer technology, and after field-trialling a modified version was produced. This version was used for the Sheffield Hallam review, referred to in LTU 20, and with very few further changes is still in use. A CD-ROM version incorporating recommendations from the Hallam report and inputs from other feedback will soon be complete.

It has been referred to as an 'oral interview'. In fact it is meant to be more than this, in two senses:

1. It involves an NS rater/interlocutor and an NNS candidate, and takes place side-by-side at a computer. It proceeds from 'acqainting' through topics that are as naturally related as possible. Assessor-training, the L1 audio help, and features of task design seek to distribute the rights and obligations of talk more evenly than the well-recorded asymmetry of OPIs, in the hope to increase validity vis-à-vis non-test encounters.

2. The tasks are interspersed with short reading passages, charts, diagrams and recorded sequences which, as far as possible, follow-on topically from the previous task. The candidate does not have to use the computer. The NS assessor types in any required input, such as the *names* referred to below. The process is simultaneously rated and navigated by the assessor's unobtrusive mouse-clicked choices. All tasks are scored tri-chotomously according to holistic criteria. The scoring is done, on completion of the task, by the assessor selecting the appropriate one of three evaluation descriptors, and clicking the button. (The evaluation descriptors are available in pop-up form, but should not need to be accessed during a test, once the assessor is familiarised and trained). The very first task, combining test registration procedures with casual discussion about names, is represented diagramatically below to illustrate these points.

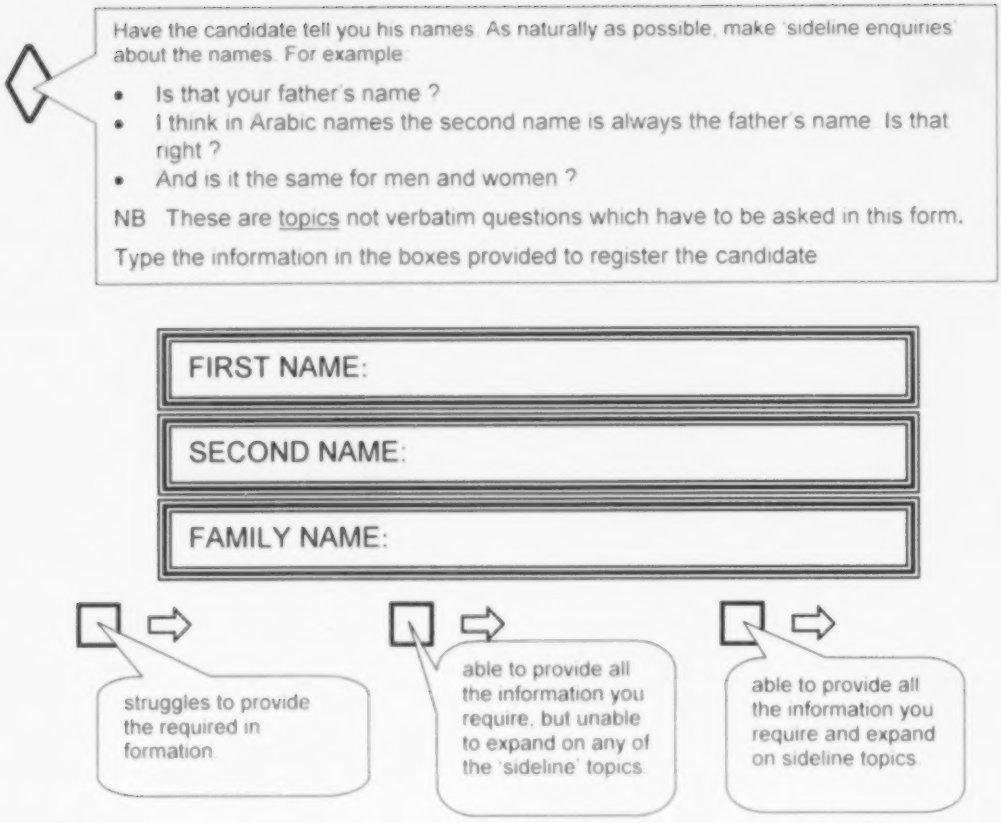
Have the candidate tell you his names. As naturally as possible, make 'sideline enquiries' about the names. For example:
• Is that your father's name ?
• I think in Arabic names the second name is always the father's name. Is that right ?
• And is it the same for men and women ?
NB These are topics not verbatim questions which have to be asked in this form.
Type the information in the boxes provided to register the candidate
FIRST NAME:
SECOND NAME:
FAMILY NAME:
struggles to provide the required in formation
able to provide all the information you require, but unable to expand on any of the 'sideline' topics
able to provide all the information you require and expand on sideline topics

Diagram Key:

- The diamond shape represents the button which activates a pop-up description of the task and how it should be conducted. (The 'speech bubble' from this diamond shows the pop-up text.)
- The three rectangles marked FIRST NAME:, SECOND NAME:, and FAMILY NAME: are name boxes where the assessor should enter the candidate's names.
- The small squares below the name boxes are buttons which activate pop-up evaluation criteria. (The 'speech bubbles' from these squares show the evaluation criteria for this task.)
- The arrow to the right represents the button which sends pre-assigned scores to the databases for the relevant constructs (in this case Listening, Speaking, and Interaction) and simultaneously selects the next task to be completed. In other words, on clicking one of these arrows, scores are invisibly collected in a database, the screen changes, and the process moves on to the next task.

The Specific Format

The idea of dealing with multiple constructs in a single event resourced from a computer occurred during the task-trialling process. Support was drawn from Nic Underhill's 1987 book *Testing Spoken Language*. It was later discovered that similar possibilities had been muted across the Atlantic, in the seminal van Lier article (1989) suggesting that *'different subparts of test batteries (Reading, Listening, Study Skills, etc) can all be included in a modular face-to-face session of no more than 30 minutes.' (van Lier, 1989)*. Van Lier had gone on to mention the possible use of computers, but had not suggested how this might be done. (Purely by chance, the average time taken to generate a Five Star profile is 29.9 minutes).

To date, three articles, plus the 'update' referred to above, have been written about this test, though nothing has been produced that would meet the standards of refereed journals. (Pollard, 1998a & b, forthcoming) The first appeared in the IATEFL Testing Newsletter in June, 1994. It was broad article covering all developments that had led to the *'imperfect prototype of a testing device that may hold potential for future development'* (Pollard, 1994: 37)

Prematurely operationalised ?

This first publication, even though un-refereed, served as a credetialling tool. Managers later showed it to directors, and though it is unlikely that they read it, it reaffirmed that language testing was a substantial discipline in its own right, and that there was a degree of professional commitment in the initiative they were funding. The conference itself also brought the project to the attention of Nic Underhill, which opened a dialogue leading to the à priori validation review. However, that degree of academic activity required more visible benefits to the shareholders than could be argued at this stage.

At the beginning of 1995 there was a further increase in the required percentages of Saudi nationals employed by large businesses, and under this pressure the parent company requested to use the test. This seemed a premature move into high-stake assessment, where life-chances were being determined, and highlighted a potential conflict between commercial issues and testing ethics – not just between the management and the tester, but *within* the tester. While the dangers were clear, excellent 'face' validity had been established using 'no-risk' volunteers, and the fears of any adverse affective reaction to the computer had been dispelled. Additionally, the results of an inter- and intra-rater reliability study, given the published limitations, suggested a further safety zone. The proposal was to test five hundred members of the population over a two-year period. This might yield more rigorous rater-reliability and predictive evidence, or at least draw valuable insights from the recipient managers who would be able to compare their new Saudis with those who had been selected using other methods.

Ageement was given, conditional that the testing would be carried out by the two assessors involved in the reliability study. There was a plan to develop an assessor/interlocutor training programme, but not until a more refined version had been developed.

By the middle of 1995 the test was being spoken of throughout the company as a considerable improvement on previous practice and this success led to it being adopted as a regular feature of the assessment centre. Further expansions were added on: a small number of military officers targetted for prestigious training colleges in the UK; employees in regional departments being targetted for ELT. Associating it with teaching in this way raised questions as to its likely use as an *attainment* measure, in spite of previous cautions. This was successfully resisted with further appeals to professionalism (quoting Alderson et al. 1995) and also by pointing out that such over-use would be likely to compromise security. Following this debate a minimum test-retest interval was established.

Seeking further evidence of validity

The compromise with commercial pressures had yielded positive feedback and, significantly, opened the door to funding the Sheffield Hallam review described in LTU 20. This was fundamentally an à priori construct validation exercise. Each task on the test putatively addressed a single skill or combination of skills, and the assessor's evaluation sent scores to the appropriate databases which drove the automatic end-of-test reports. It was essential that these tasks did, in fact, address the relevant skills, and this was the main issue under scrutiny.

While the above interactions with management forces had sustained the project this far, the search of the literature that had accompanied the development had raised many

uncertainties. For example, since the 'core' event was a dyadic NS-NNS interaction, this related to the twin questions raised towards the end of the eighties:

- To what extent were the interactional processes involved similar to normal conversation ?
- To what extent could these processes be seen as skills attributable to the NNS which generalised to oral interactions in the target use domain ?

(*Paraphrased from van Lier, 1989:489 & 501*)

In applied linguistics evidence has been mounting that interaction is somehow fundamental to second language use and its inseparable correlate, second language acquisition. In addition to a renewed interest in Vygotskian 'scaffolding' in the communication and learning processes this has recently been examined in a number of related branches of research, including:

- Second Language Acquisition [SLA] e.g. Færch & Kasper, 1984; Kramsch, 1986 ; Ellis, 1991
- Second Language Classroom Research e.g. Chaudron, 1988; Long, 1983; Pica, et al 1989-1996; Johnson, 1995
- Conversation Analysis [CA] e.g. Sacks, Schegloff, et al 1974-1995; Atkinson & Heritage, 1984; Jacoby & Ochs, 1995; Eggins & Slade, 1997
- Second Language Testing Research e.g. Shohamy, 1983-93; van Lier, 1989; Ross, 1992 & 1994; Ross & Berwick, 1992; Young & Milanovic, 1992; Zeungler, 1993; Young, 1994; Wigglesworth, 1994; Lazaraton, 1992 & 1996

Idealised models of language seemed to underpin the tests previously used in this context. In more realistic models of second language performance, it seemed that the constructs reported should include ability in the dynamic processes of real encounters. The strongest expression of this comes from CA, claiming that interaction is 'the primordial locus for the development of language, culture, and sense-making' (Jacoby & Ochs, 1994: 187)

Moreover, this may be particularly pertinent in this cultural context. One of the population characteristics seems to be stronger SL oral than literary abilities – which may reflect the balance in L1. Comparisons of Arab and North American nonverbal communication reveals greater engagement in the former – closer proximity, increased gesticulation, touch and eye-contact (Watson & Graves, 1970). Our own observations seem to confirm this view. It has also been noted that, in post-test discussions, the notion of co-construction met with ready acceptance.

On this basis, and with reference to applied research on classroom interaction, a construct definition of interaction had been included in the original test specifications, though it was restricted to the more extensively documented features of oral negotiation:

'Interaction : a learner's ability to facilitate participation in a one-to-one discussion through the employment of negotiation devices such as confirming understanding, requesting repetition and seeking clarification.'

This construct had not been examined in the Sheffield Hallam review, as it was thought to overlap with the other 'core skills' of Listening and Speaking, and the methodology 'theoretically demanded independence between skills'. There was also the practical consideration that in phases 1 and 2 of the review the panel were only examining the 'cold' tasks. In phase 3 they would view video recordings of live tests to provide a first structured comparison between 'Five Star' and ESU framework levels. This opportunity was seized to conduct a 'sub-survey' exploring the following questions:

1 Was a construct domain of *Interaction* salient to informed observers?

2 Were there were any clear patterns in the frequency and density of interactional features ?

3 Did interactive behaviours contribute significantly to the completion of the test tasks ?

First the panel members were familiarised with a key, and then completed an observation matrix while watching the videos (See appendices 1 and 2).

The first two questions were affirmed by the data, and the concensus of the exercise was that *'the Five Star test can be seen centrally as a test of direct interaction between interlocutor and participant'*. (Underhill, 1996). This sub-survey also identifying the task types that coincided with most interaction, and revealed that a great deal took place outside 'task boundaries'. For example, although Arabic recorded instructions were used for the earlier, less challenging tasks in order to eliminate contamination, later tasks relied on NS assessor explanations in L2. These parts of the process revealed a high density of interactional features. Interestlingly, if we borrow the 'academic task structure' and 'social structure' distinction made by Karen Johnson (1995) vis-à-vis SL classrooms, these phases of the process represent some of the most authentic uses of target language. It is also revealing that 'receiving explanations' does not seem to be included in any functions-oriented tests, yet appears to be a prevalent event in our target situations of NNSs entering new jobs or technical and administrative courses. In the new version of the test now being developed into CD-ROM, tasks where this occur have been increased, and the evaluation split so that an interactional and listening score is derived for the successfully negotiated explanation in the target language.

Conclusion

The direction of the enquiry outlined here demonstrates that it is difficult to develop tests with reference to models that 'reflect more closely the beliefs of the time' (Alderson & Clapham, 1992: 149) without getting involved in issues that are poorly understood and little researched. It may be that what takes place in this particular assessment process does more closely resemble 'normal conversation' than more traditional OPI formats. It may also be that it captures essential interactive competencies (Kramsch,1986) and that these may underlie performance in post-test situations, justifying van Lier's speculations and fulfilling Messick's generalisability criterion (Messick, 1994). However, it would be both unsatisfactory and wrong to assume any of this.

From the above experience it seems clear that new developments in testing are not only triggered 'top down' from theory and research, however ideally desirable that may be. Test developers working outside the framework of academic or institutional operations are subject to forces that are just as likely to prise open Pandora's box as those whose main concerns are theory and research. This surely argues for a role of *research-in-development*. It confirms that test development needs 'insights from applied linguistics', but suggests that language testers can 'serve applied linguistic theory' in more expansive areas than the 'practicalities of test design and administration' (Alderson and Clapham, 1992:165).

[Article 2. The follow-up article will explain how opportunities to use data from such a test-in-development led to more substantial research designs which could be exucuted in-development to improve assessor-training and reporting descriptors.]

References:

Alderson J C & Clapham C. 1992. *Applied Linguistics and Language Testing: A Case study of the ELTS Test.* Applied Linguistics 13/2:149-167.

Alderson J C, Clapham C and Wall D. 1995. *Language Test Construction and Evaluation.* CUP.

Alderson J C, Krahnke K J & Stansfield C W (Eds). 1987. Reviews of English Language Proficiency Tests.TESL, Washington DC.

Atkinson J M and Heritage J. 1984. *Structures of Social Action: Studies in Conversation Analysis.* CUP.

Chaudron C. 1988. *Second language classrooms: research on teaching and learning.* CUP.
Eggins S & Slade D. 1997. *Analysing Casual Conversation.* Cassell.

Jacoby S & Ochs E, 1995. *Co-construction: an introduction.* In Jacoby & Ochs (eds) *Special issue on co-construction.* Research on Language and Social Interaction 28/3: 171-83

Johnson K E, 1995 *Understanding Communication in Second Language Classrooms.* Cambridge University Press.

Lazaraton A, 1992. *The structural organisation of a language interview: a conversation analytic perspective.* System: 20: 373-86

Lazaraton A, 1996. *Interlocutor support in oral proficiency interviews : the case of CASE.* Language Testing 13/2: 151-172

Long M H,1983. *Native speaker / non-native conversation and the negotiation of comprehensible input.* Applied Linguistics speaker 4/2: 126-41

Ellis R, 1991. *The Interaction Hypothesis: A Critical Evaluation.* In Sadtono, E. (Ed) *Language Acquisition and the Second Language Classroom.*

Færch C & Kasper G, 1984. *Two ways of defining communication strategies.* Language Learning 34/1: 45-63.

Kramsch C, 1986. *From Language Proficiency to Interactional Competence.* The Modern Language Journal, 70/iv

McNamara T F, 1995. *Modelling Performance: Opening Pandora's Box.* Applied Linguistics 16/2:159-179

McNamara T F, 1996. *Measuring Second Language Performance.* Longman

Messick S, 1994. *The interplay of evidence and consequences in the validation of performance assessments.* Educational Researcher 23/2: 13-23.

Pica T, 1994. *Review Article. Research on Negotiation: What Does It Reveal About Second-Language Learning Conditions, Processes and Outcomes ?* Language Learning 44/3.

Pollard J D E, 1994. *Paper on Proficiency Testing.* IATEF Testing Newsletter July, 1994.

Pollard J D E and Underhill N, 1996. *Developing and Researching Validity for a Computer-Resourced Proficiency Interview Test.* Language Testing Update 20: 49-52, ILTA.

Pollard J D E, 1997: *Saudi Development and Training's 'Five Star' Proficiency Test Project.* In Coombe C A (Ed): *Current Trends in English Language Testing : Conference Proceedings for CTELT 1997 and 1998.* TESOL Arabia, Al-Ain, United Arab Emirates.

Pollard J D E, 1998: *The Influence of Assessor Training on Rater-as-Interlocutor Behaviour During a Computer-Resourced Oral Proficiency Interview-cum-Discussion (OPI/D) known as the 'Five Star Test'.* In Coombe C A (Ed): *Current Trends in English Language Testing : Conference Proceedings for CTELT 1997 and 1998.* TESOL Arabia, Al-Ain, United Arab Emirates.

Ross S and Berwick R, 1990. *The discourse of accommodation in oral proficiency examinations.* SSLA: 14:159-76. Anthology Series SEAMEO Regoinal Language Centre (RELC), Singapore.

Ross S, 1992. *Accommodative questions in oral proficiency interviews.* Language Testing 9/2: 173-186

Ross, S (1994). *Formulaic speech in language proficiency interviews.* Paper presented at the annual conference of the American Association for Applied Linguists, Baltimore MD, March.

Schegloff E A, 1995. *Discourse as an Interactional Achievement III: The Omnirelevance of Action.* In Jacoby S & Ochs E (eds) *Special issue on co-construction.* Research on Language and Social Interaction 28/3:185-211

Shohamy E (1983). *The stability of oral proficiency assessment on the oral interview testing procedures*. Language Learning 33/4: 527-40

Shohamy E (1988). *A proposed framework for testing the oral language of second/foreign language learners*. SSLA 10/2: 165-79

Underhill N (1987) *Testing Spoken Language*. Cambridge University Press

Underhill N (1997) *Unpublished research report*

Van Lier L, 1989. *Reeling, Writhing, Drawling, Stretching, and Fainting in Coils: Oral Proficiency Interviews as Conversation*. TESOL Quarterly, 23/3

Watson O M & Graves T D. 1966. *Quantitative research in proxemic behaviour*. American Anthropologist 68:970-985

Watson O M, 1970. *Proxemic Behaviour: A Cross-Cultural Study*. Mouton

Wigglesworth G, 1993. *Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction*. Language Testing,10/3: 305-335 Wigglesworth G (1994). *An investigation of planning time and proficiency level on oral test discourse*. LT 14/1: 84-106.

Young R and Milanovic M, 1992. *Discourse variation in oral proficiency interviews*. SSLA 14:403-24

Young R, 1994. *Conversational styles in language proficiency interviews*. Language Learning 45/1: 3-42

Zuengler J, 1993. *Encouraging learners' conversational participation: the effect of content knowledge*. Language Learning 43: 403-32

APPENDIX

K E Y

- ◆ **CONFIRMS UNDERSTANDING**
- ◆ **SEEKS CONFIRMATION**
- ■ **SEEKS CLARIFICATION**
- ■ **INDICATES NEED FOR CLARIFICATION**
- ● **CONFIRMS OWN PREVIOUS TURN**
- ★ **RE-FORMS OWN PREVIOUS TURN**

◆ CONFIRMS UNDERSTANDING 1 Offers appropriate response 2 Says "I understand" or = 3 Says "Yeh", "Yeah", "u-uh", etc. 4 Agrees 5 Disagrees 6 Laughs (appropriately)	◆ SEEKS CONFIRMATION 7 Asks "Do you mean + *item* ?" or = 8 Repeats interviewer words/segments with questioning intonation 9 Refers with deixis - "You ?", "Me ?", "Here ?", "This ?"
■ SEEKS CLARIFICATION 10 Says "I don't understand" 11 Says "I'm sorry ?", "Excuse me ?" or = 12 Says "Please repeat" or = 13 Repeats part of words/segments from interviewer's turn with obvious uncertainty	■ IMPLIES NEED FOR CLARIFICATION 14 Fails to respond / extended silence 15 Responds with disfluencies such as "errrrr....", "ermmm....."
● CONFIRMS OWN PREVIOUS TURN 16 Says "Yes" or = 17 Says "That's right" or = 18 "Yes, + *item* " 19 Repeats *item*	★ RE-FORMS OWN PREVIOUS TURN 20 Says "No, what I meant was (or =) + repeats/rephrases *item*" 21 Rephrases *item*

INTERACTION - SAMPLE OF COMPLETED OBSERVATION MATRIX (Numbers in cells represent the instances of occurrence recorded by observers:

Identify the features of verbal interaction used by the learner. For each occurrence of a listed feature, put a tick in the cell.

CANDIDATE: **Fahad Al-Radass**	VERBAL INTERACTION (facilitating self-understanding)				**VERBAL INTERACTION (facilitating other-understanding)**	
TASK	1 CONFIRMS UNDERSTANDING	2 SEEKS CONFIRMATION	3 SEEKS CLARIFICATION	4 IMPLIES NEED FOR CLARIFICATION	5 CONFIRMS OWN PREVIOUS TURN	6 RE-FORMS OWN PREVIOUS TURN
4 Names	9	2	4	4	4	0
11 Numeracy	8	5	0	2	1	9
13 Al Harbis	8	1	0	0	4	3
15 Student Grades I	11	2	0	0	1	3
19 Student Grades II	6	2	0	5	2	1
23 Vehicles	3	4	6	3	1	0
24 Footballers	2	2	1	2	0	3
26 Kettle	1	0	0	0	0	0
28 Signs I	5	2	2	0	3	0
29 Fridge	8	3	2	5	1	0
50 Signs II	3	1	0	0	0	3
55 Kuwait City	5	1	1	4	3	0
62 Car ownership	8	1	0	0	4	4

BEST COPY AVAILABLE

Research and development: a complex relationship - Part II

This is the second of two reports. The first (LTU 24, 1998: 46-57) described a situation where commercial constraints necessitated the integration of research, development and implementation - roles typically segregated in language testing. The increasing interest in qualitative research in relation to OPI-like events offers opportunities for improving assessment procedures under these conditions, and perhaps informing on issues of wider concern. This follow-up suggests the benefits of such research-in-development to a specific test project and findings that might be of interest to others involved in live language assessments.

There still seems to be inadequate qualitative evidence to meaningfully evaluate OPI-like events as samples of 'interactional competence' (Kramsch, 1985) or dismiss them as inauthentic (Young & Milanovic, 1992; Young and He, 1998, et al). State-of-the-art comments on the application of insights and practices from Conversation Analysis (CA) urge the involvement of more test developers in this process (Lazaraton, 1992:383; 1996:167). There is a need to assess the impact of varied test formats and task features on interaction (Chalhoub-Deville, 1995:17) and to explore how these events vary in different cultural settings. The former is emphasised by the development of assessments, like this Arabian test, with a dyadic *human-plus-computer* configuration (e.g. Malabonga, 1998: 29); the issue of cultural influences is widely referred to and recorded in the literature (e.g. Bachman, 1990: 187 and 273; Porter, 1991; White, 1998).

The consensus model of second language performance now includes the interactive elements associated with strategy and co-construction in communication (van Lier, 1989; He and Young, 1998:3). Formats and content specifications are being modified to incorporate these additions (Pollitt, et al, 1995:18; Kromos, 1998), and new notions of *task* are likely to emerge (Fulcher, 1994). Without a clear relationship through all of these elements to the scoring, reporting, and use of assessments, it is difficult to identify the multiple sources of data that many recommend for validation evidence (Shohamy, 1994). We are still fundamentally deficient in our knowledge of *how discourse features are ... reflected in ratings* (Upshur and Turner 1999).

In this particular initiative the absence of extensive systematically-gathered data on the *target language use domain* (Bachman and Palmer, 1996: 18) had motivated an *a priori* 'expert panel' design for initial evidence of construct validity (LTU 20:49-52). However, for the facets of construct validity that reside in *generalisability* (Bachman, 1990:256; Messick, 1994; McNamara, 1996:15) the developer had to rely on indirect evidence that conversational interaction was 'bedrock' to the construct of interaction being assessed (Heritage, 1989:33-34; van Lier, 1989:493-6; Shegloff, 1995).

The design of the test was an attempt to maximise this conversational style of interaction through features of test design such as the computer program that mediated the event, and assessor training. The former included:

- A succession of holistic evaluations incrementing to a final score, reducing the processing and memory burden on the NS interlocutor.
- The positioning of these evaluations at junctures that *could* allow the NS interlocutor to avoid sudden pre-closure and closure moves once the assessor has 'got what he wants' from the candidate.

It was hoped that these features would also minimise distractions such as having to refer to evaluation criteria during the event. (see, for example, Lazaraton, 1992:378)

- Topics of personal and local interest
- Topic succession planning

These were intended to endow the candidate with 'knower' status to authenticate NS curiosity (Woken & Swales, 1989; Zeungler, 1989:238, 1993) and to minimise the 'topic-hopping' that confirms many OPIs as interviews rather than conversations.

It was not anticipated that these features would work without assessor training that focussed on NS interlocutor support. During the early trialling of the assessment process this was conducted by the developer on a one-to-one basis over a number of days. It consisted of:

1 Concept familiarisation (principles and purposes)
2 Observations of the test being administered (with feedback)
3 Technology familiarisation ('knobs and buttons')
4 Observations of video recordings of sample tests (with feedback)
5 Role-play, (alternately playing the parts of assessor and candidate)
6 Assessments of real but 'low-stakes' candidates while being observed
7 Video-recording test performance (for feedback). In this phase the trainee was equipped to make a number of video recordings of himself conducting actual tests, was asked to return samples and (if willing) signed an agreement that the videos would be used for research.

In 1995, just after the 'prototype' had been developed, managers working in a separate location urgently requested assessments when no trained assessors were available. Cautions were overruled, and the above training procedures by-passed. Two new members of the ELT staff were provided with the test program. They familiarised themselves with the tasks and agreed to video some of their assessments for feedback, but received no other training before the assessments were conducted.

Though potentially hazardous, this emphasises the point in the previous report about the potential benefits of pragmatic management in test development.

NS ELT staff viewed the videos and acted as informants on the behaviours of the assessors. The dominant theme in their informally recorded impressions was the different *rapport* between assessor and candidate, manifested through 'tone of voice' (described as 'formal' / 'informal'; 'friendly / 'distant'; 'reassuring' / 'unconcerned') and 'attentiveness and interest' ('isn't interested'/ 'takes an interest'; 'doesn't really care'/ 'shows understanding'). One remark referred to 'fidgeting' behaviour, and another noted the different spacial distance between the participants.

Research links *affective schemata* and second language performance (Tarone, 1980, 1982 and 1983; Faerch & Kasper, 1984; Eisenstein & Starbuck, 1989; Bachman and Palmer, 1996:66; McNamara 1996: 73). It is also widespread practice to preface live interactive tests with a 'warm-up' phase to establish a friendly and conducive atmosphere (Lazaraton, 1992:382; Young, 1995:18). However, with the exception of a small part of a study relating to the Australian OET (McNamara & Lumley, 1997), there appear to be no systematic attempts in language testing research to define rapport, or explore its impact on assessments. As we have moved nearer to a co-constructed model of second language performance such an enquiry seems appropriate, if not overdue.

A structured data-gathering session was planned, using audio recordings and Likert-scaled feedback forms. Since the commercial constraints ruled out research *per se*, this was incorporated in an assessor training programme as an awareness-raising activity. In small 'syndicate' groups, the trainees listened to sections of assessments audio-recorded from videos, marking points on scales for a number of features associated with rapport. The results were used to select samples that would be qualitatively examined.

In this process an attempt is being made to follow the detailed and iterative scrutiny which CA applies to 'naturally occurring' samples of conversation. The methodology being adopted is that of Grounded Theory (Strauss and Corbin, 1990; Cresswell, 1998). However, as 'naturally occurring' and 'testing event' are contradictory, it might be more accurate to describe this as *Interactive Language Test Analysis*.

Both CA and Grounded Theory belong to the qualitative research tradition, and are essentially *discovery* techniques (Strauss & Corbin, 1990:23; Silverman D,1993; Cresswell, 1998: 150). The aim is to construct rather than test theories (Hopper, 1989:52).

The following observations focus on samples of assessments that ranked at the top and bottom of the 'rapport table'. For the purposes of this discussion, *rapport* refers not only to the general ambience, but to detailed evidence of the interest, attention, relevance, good faith, and interlocutor support that is associated with conversational norms (Grice, 1975:46; Heritage, 1989:28; Tannen D, 1989:89; Eggins S & Slade D, 1997:36). The search has been for characteristics of *reactive* and *mutual contingency* rather than the *asymmetric contingency* of interviews, as described by Jones and Gerard and subsequently adopted by van Lier (1989) Young (1992) and others. The features discussed range from aspects of discourse which extend over stretches of talk (e.g. topic; rhetorical structure) to micro-focussed turn-constructional features (e.g. reformulations; news-marking; ratification devices; backchannels; gaze; inter-turn space). Excerpts below illustrate how co-occurrences of these features are patterning in the earliest stages of the enquiry.

KEY TO TRANSCRIPTION CONVENTIONS

The transcription conventions used above are those developed by Gail Jefferson (1983) and extended in Atkinson and Heritage (1984). For a full review of these conventions, see Ten Have, 1999:75-98. The only departure from these conventions is in the use of underscoring and blocking for marking onset, maintenance, and withdrawal of mutual eye-contact or 'gaze'.

1 2 3 4	The serial numbers down the left-most column are use for the identification of specific parts of the transcription.
→	The right-pointing arrows in the second column from the left indicate features referred to in the analysis text.
C A	Capital C and A in the third column from the left refer to the turns of (C) the candidate, and (A) the assessor.
(1.4), (2.4), etc	Whole and decimal numbers inserted between parenthesis within and between turns in the transcription represent pauses in seconds and hundredths of seconds, as measured using a digital stopwatch.
( ) – (. . . .)	One or more periods between parenthesis indicate a range of perceptible pauses too small to measure using a digital stop-watch. A greater number of periods indicates a longer pause.
°what ?°	Small single superscripted circles enclose stretches of talk in quiet or softly spoken voice.
°°yes°°	Small double superscripted circles enclose stretches of talk in very quiet or softly spoken voice.
↑raining	The bold upward-pointing arrows indicate that the following element is produced with increased amplitude.
↓father	The plain text downward-pointing arrows indicate that the following element is produced with decreased amplitude.
C man= A =yeh= C =his	Equals signs indicate intra- and inter-turn contiguity, or absence of pauses.
[. . .]	Italic parenthesis are used to enclose relevant nonvocal behaviour such as gesture – for example: *[moves hand down piano-playing-like finger movements]*
ra[in ? [yes	Square parenthesis indicate the onset of interruption or simultaneous talk.
what do	Underscoring followed by blocked text. The underscoring marks the transcribed section of talk accompanying the onset of eye-contact ('gaze') by the speaker.
from Abha	Underscoring following blocked text. The underscoring marks the transcribed section of talk accompanying the withdrawal of eye-contact ('gaze') by the speaker.
( )Okay( )	Blocked text indicates that mutual eye contact is maintained during the transcribed section of talk that has been blocked.

BEST COPY AVAILABLE

NS Reformulations (Low Rapport assessment)

Reformulations appear to be less evident in samples given a 'poor rapport' rating. Transcript Extract 1.1 is an example, showing a *reduction* of context from the assessor's first turn (line 1) to 'it's raining' in line 4, and then an exact repetition of this reduced form (line 7). The first reformulation (line 10) merely fronts the key item 'raining', followed by a third repetition.

Transcript Extract 1.1 [Candidate:20/001]

1		A	so it's raining today.
2			(1.4)
3		C	°what ?°
4	→	A	it's ↑raining (..) °come forward° (.....) You like the rain ?
5			(2.4)
6		C	what ?
7	→	A	it's raining.
8			(0.4)
9		C	ainy ?
10	→	A	↑raining (.) it's raining
11		C	raining ?
12		A	mmm *[moves hand down piano-playing-like finger movements]*

NS Topic selection and management (Low Rapport assessment)

Saudi Arabia is one of the driest countries in the world. While it was raining on this occasion, the topic of changeable weather introduced in Transcript Extract 1.1 is not generally used for phatic 'small talk', as in some cultures. Furthermore, the test room had no windows to give the topic visibility. Although this may reflect an overall lack of cultural awareness on the part of the assessor, what follows is of more concern. In Extract 1.2 the NS pursues a rhetorical goal-orientation which prevents mutual or reactive patriciption or co-construction, as shown in this analysis: The candidate's affirmative 'yeses' (line 16) overlapping with the assessor's turn (line 15) 'do you like the rain ?' cannot be a response to the question. This is confirmed by the candidate echoing the repeated question (line 18, Extract 1.2:) – a strategy used earlier when understanding failed (*iany* – 1.1: line 9). They probably refer to *'it's raining'*, repeated and demonstrated through lines 1-10 (Extract 1.1). Following this, a further assessor repetition is overlapped with a 'yes' from the candidate (line 21) who, at this point, identifies the topic. This is confirmed in the 'yeses' repeated with increased volume and stress (line 22). These seem to combine topic ratification and confirmation signalling, indicating *news-worthiness*. A preferred conversational response would be a newsmark invitation to take the floor and expand (Heritage, 1989:30). Instead, the assessor laughs at the induced contradiction, pre-closes with ↑ok↓ay, (line 27) and moves on to the next task.

Transcript Extract 1.2 [Candidate:20/001]

12		A	mmm *[moves hand down piano-playing-like finger movements]*
13			ah yes ( ) by the mm=water
14		C	mmm do you like the ra[in ?
15		A	[yes=yes
16	→	C	°do you like it ?°
17		A	(1.02)
18			er (.) d'yer luk et
19		C A	↑do you like it? Do[you like the rain?
20		C	[↑yes
21	→	C	↑yes (1.02) ↑yes
22	→	A	do you like to drive in the rain ?
23		C	°no°
24		A	HA[H (...)
25		C	[no
26		A	↑ok↓ay (.) let's start the next task
27	→		

Reformulation (High-rapport assessment)

In line 64 of Extract 2.1 confirmation is sought about the Arabic naming system. The reformulation of this complex proposition involves both the fronting *'In Arabic names'* and the replacement of the elliptical *'the second name'* with *'any man . . . his second name'* (also an instance of foreigner talk). This reformulation is negotiated over an extended turn, serviced by the candidate with the backchannelling confirmation token (*yeh* – line 70), indicating close attention associated with reactive contingency. The topic-related question in line 73 (*Is that the same for girls ?*) is also met with the clarification request (*sorry* - line 74). The assessor offers this again, un-reformulated, but retains the turn by repeated *if*s (line 75) ahead of exemplifying and personalising the question to the candidate. This turn is also an extended negotiated one with candidate backchannelling (line 76). In both instances, there is a clear co-construction of meaning, and evidence of an ability and willingness on the part of the NS to negotiate rather than assume shared meaning.

NS Topic selection and management through reformulation, news-marking and ratification (High-rapport assessment).

In line 71 of Transcript Extract 2.1, the assessor newsmarks the information elicited about the local naming system (a topic he nominated). The topic is then pursued on the basis that the candidate is the knower. The result is the long productive string : *the second name is er father* (line 79). No utterance of comparable length or coherence was elicited in the other sample - which means that one candidate was fulfilling the task criterion (of being able to expand on the topic) while the other was not. Although there is no way of confirming this at this stage, it is likely that the different NS interlocutor behaviour was reflected in the candidates' scores.

Transcript Extract 2.1 [Candidate:25/001]

60		A	D Shadeed
61		C	°Shadeed° °°yes°°
62		A	[Is that your father's name ?
63		C	yeh this my °↓father° =
64	→	A	=yeah. I think in Arabic names (.) the second name is always the father's name isn't
65			it ?
66		C	er (.) sorry
67		A	yeh in Arabic names the e (.) any man=
68		C	=yeh=
69		A	=his second name is (.) same as his father=
70		C	=°yeh yeh °=
71	→	A	=yeh=
72		C	°that's right°
73		A	is that the same for girls ?
74	→	C	(.....) Sorry
75	→	A	is that the same for girls? If(.)if(.)if you have a sister, for example
76	→	C	°yeh°
77	→	A	erm (...) an her second name is the same as your father also in
78			Arabic names
79		C	(.) errrrrr (.) the second name is er father
80	→	A	always ?
81		C	always, yeh
82		A	[yeh man or woman, girl or boy
83		C	yyyeh tht's right
84		A	[yeh yeh yeh okay

NS Discourse strategy and inter-turn space (Low-rapport assessment)

The 'low-rapport' assessor appears to have a strategy of asking serial questions. This is illustrated in Extract 1.3 (lines 54, 58, 62, 65, 67 and 70). Though thematically related, these appear to lack the topic contingency of natural conversation. This section of talk is also characterised by lengthier 'unmanaged' inter-turn spaces (averaging 1.6 secs.) than those occurring in comparable stretches of 'high-rapport' assessments, and an absence of candidate backchannelling. 'High-rapport' assessments do not contain serialised questions in this way, but where questions are used the inter-turn spaces average 1.00 sec., more in line with inter-utterence space in CA data (e.g. Jefferson, 1989:167-192). There is evidence of backchannelling as remarked above. More noticeable, however, is the high frequency of contiguous turns associating with high-rapport. *A brief glance at* Transcript Extract 2.1 shows how frequent contiguous turns (marked with an = sign) are in this stretch of dialogue. There are no contiguous turns in otherwise equivalent sections of the 'low rapport' assessments – for example, in Transcript Extracts 1.1-3

Transcript Extract 1.3 [Candidate:10/001]

54	→	I	where does the Al-Shehri family come from ?
55		C	(1.05) from er ( ) Nammas
56		I	Nammas
57		C	yeh
58	→	I	where's that ?
59		C	from Abha
60		I	↑aah – near Abha I was there ( ) yesterday
61		C	yesterday ( ) Abha
62	→	I	and where do you live ?
63			(1.92)
64		C	I live er Musaillatt
65	→	I	where's that ?
66		C	Musaillatt (1.85) east er Riyadh ( ) ↑or south Riyadh
67	→	I	°hmm° (0.4) ↑now (0.2) ↑far (0.2) from Riyadh °is it° ?
68			(2.03)
69		C	twenty (1.70) kilo°meters° (0.4) twenty kilometers
70	→	I	twenty kilometers ( ) ok↓ay (0.7) so (0.4) what do you do there ? ( ) do
71			you work ?
72			(1.22)
73		C	no I work in er ( ) airport er ( ) King Khalid airport
74			( ) before that I work in a bank

NS-initiated Eye-Contact or 'Gaze' (High- versus low-rapport assessments)

In Transcript Extract 1.3 above this behaviour is marked at the point of NS gaze onset and withdrawal. Instead of the CA convention of overscoring stretches of discourse with mutual gaze, **blocking** has been used here, and the periods of withdrawal referred to are underscored for easy reference. The onset and withdrawal points are the transitions between these markings (e.g. onset: what do you do; withdrawal: **from** Abha). As data on this feature accumulates, it invites a number of inter- and intra-assessment comparisons. Instances of gaze appear to be more frequent and of longer duration in high-rapport assessments. The positional occurrence is also of interest. For example, in 'low-rapport' assessments, there is a consistent NS withdrawal of gaze just ahead of each serial question. This also coincides with consistent NS face and eye roaming (lines 54, 62, 67 and 70). One interpretation would be that the NS assessor is using this space to think of a next question to ask. A definition may emerge through this data which directly associates gaze behaviour with mutual and reactive contingency.

BEST COPY AVAILABLE

Finally, even in the 'low rapport' assessment, the pattern attributed to the NS strategy of serial questioning and associated inter-turn space does not prevail throughout the entire event. In Extract 4 below we observe a quickening of pace and other indicators of co-construction in one section of the talk.

This extract is also interesting since it appears to reveal a lost opportunity. More data will be necessary to collaborate this, but it seems as though the closer contingency (evidenced in NS gaze, NS newsmarks, NNS backchannels, and contiguous turns, in lines 53-55) results in the candidate being more productive (perhaps, less careful) as revealed in the extensive turn in lines 56-57. The turn-retaining *and er* (line 59) indicates a willingness to expand even further on the ratified topic. In fact the candidate does so, but is met with gaze withdrawal and backchannel dysfluencies in diminishing pitch and volume which serve as pre-closures to the topic change in line 67. In fact during this lengthy pre-closure, the candidate's turns fade in volume as, in all probability, he detects the withdrawal of interest. This accounts for the inability to transcribe sections of this talk in lines 59-66.

Transcript Extract 1.4 [Candidate:10/001]

48		I	who do you speak English to ?
49			(1.92)
50		C	er (0.2) many people ( ) American ( ) Saudi
51	→	I	↑and Saudi ?
52		C	yeh (.) too much Saudi
53	→	I	you speak English with the Saudis ?=
54		C	=yeh=
55	→	I	=really ?=
56		C	=yeh oh ( ) er ( ) ( ) yeh because er ( ) my work in communication
57			training, typing er [ )
58		I	[mm
59		C	and er (° °) (0.5) (° °)
60		I	security ?
61		C	no (1.06) er planner ([ )
62		I	[mm ↓hmm
63		C	(°n °)
64		I	(3.19)
65		C	°mm hmm°
66		I	(° °)
67		C	↑right ( ) what's your application number please ?

This is a brief illustrative sketch of some features of interactional behaviour that are beginning to pattern in this sample of data. The following conclusions speculate on how a fuller sequel of this process might usefully inform the construct of proficiency and theoretical aspects of validity, as well as providing practical inputs to test design and assessor training.

In the absence of stable or empirically established theoretical models (Shohamy, 1988:167; Chalhoub-Deville, 1997:4), this type of analysis, combined with the flexibility of the computer's 'open architecture', seems to offer a more progressive approach to test development than the 'recency avoidance' attributed to the pace of change in the theoretical arena (Skehan, 1990:4). It is hoped that the insights which emerge will lead to improvements in task, evaluation criterion, reporting devices, assessor training, and, as a *unitary concept* partly constituted of these facets, construct validity (Bachman, 1990: 236 et seq).

Proficiency and Construct Validity

It has been pointed out that construct validity applies not to tests, but to the uses made of them. (Bachman, 1990:246, exemplifies this referring to multiple-choice vocabulary test). This implies describing all and only the relevant conditions under which the test behaviours are produced. A better understanding of *rapport*, subsuming 'recipiency' and 'co-participation'

BEST COPY AVAILABLE

(Atkinson & Heritage, 1984:224; Goodwin, 1984:236), seems to be a relevant factor in the performance conditions reported in this test. An anecdote serves to illustrate. The test is used as part of an 'assessment centre' for recruiting Saudis to a British company. The process includes other ability tests and a panel interview focussing on personal and professional considerations, conducted partly in English. Unfavourable comments were made by a NS panel member about one candidate's language ability – even though this was not the purpose of the interview. The inclusion in the English proficiency report of phrases like *'in an informal one-to-one situation with an experienced and supportive native speaker'* helped to clarify aspects of the interview that were likely to account for the 'poorer' performance. This reinforced the credibility of the English assessment, and permitted a more balanced evaluation of the candidate's language proficiency. He was recruited and, though not an outward-going individual (hence, perhaps, the different impressions across the two events), co-workers now speak highly of his performance. This recalls recent comments on the *consequential* aspects of validation (Cumming, 1995:6).

Assessor Training

There has been a call for different, more extensive, and recurrent training of direct oral language assessors (Underhill, 1987; van Lier, 1989; McNamara and Lumley, 1997). Some urge more rigour in the selection and authorisation of assessors, and increased post-training monitoring (McNamara, 1996: 238) Specifically, there have been recent calls for training of NS *as interlocutors* as well as *assessors* (McNamara and Lumley, 1997:154). An improved understanding of interlocutor behaviours emerging from this enquiry may help to provide more focussed examples of assessor performance in training sessions.

A number of studies have questioned the impact of assessor training from the perspective of judgement reliability (Barnwell, 1989; Ross, 1992; Ross and Berwick, 1992; Wigglesworth, 1993; McNamara, 1996: 235; Lynch and McNamara, 1998). The impact of assessor training in *interactional awareness* remains to be examined. However, it is encouraging that sections of talk that vary in their similarity to conversation can be detected in 'low rapport' assessments. It is also promising that potential opportunities to be more faithful to non-test encounters occur in these samples.

In the 'it's raining' extract, topic selection speaks for itself. However, there is an interplay between topic selection and topic development. Compared to natural conversations live language assessment events are agendered and pre-determined in some structural aspects. In both, however, participants depend upon repertoires of rhetorical structure, and it is inevitable that assessors (who may conduct hundreds of assessments) will develop such repertoires. In non-English L1 environments there is a need for these repertoires to cater for culture-specific expectations. By detecting repertoires which are likely to be in conflict with these expectations, such as the 'raining' topic, we may be able to pre-empt their deployment in future tests.

Research-in-development.

The value of qualitative research is becoming apparent, and this report suggests that it might be an accessible and productive form of enquiry for researcher-developers. However, it may also provide opportunities to compliment work where the preference is to construct à priori theories for experimental designs (Hopper, 1989; Lazaraton, 1999). Much recent OPI research seems to be of this latter type, often citing van Lier on the lack of similarity between OPI-like tests and 'real' conversations (van Lier, 1989:505), and then selecting features from CA to construct hypothesis-testing designs
(Young and Milanovic, 1992; Young, 1995; Johnson and Tyler, 1998). While all types of research have their contribution to make, there does seem a slight prematurity to this approach. Van Lier's central contention was that OPIs *'could be made more like real conversations'*, and as a first step towards this, directed researchers to the qualitative discovery approaches of CA. While there have been such analyses of OPIs based on long-established tests (Brown, 1998), tests for specific course placement (Lazaraton, 1992), and

vocational tests (McNamara and Lumley, 1997), this research is still 'in its infancy' (Lazaraton, 1996a:156), and van Lier's key practical proposition remains largely unimplemented.

Upshur and Turner's recent comments on the lack of a theory linking discourse in tests with scoring mechanisms is echoed in a recent review of Alderson and Clapham's work (Chapelle and Douglas, 1999:116).

We still rely on unfounded notions of relative difficulty in second language performance. The relationship between knowledge and performance is unclear (McNamara, 1996: 183), and communicative functions (such as can be found in the work of Wilkins, 1976, and Munby, 1978) are characterised by *differing levels of abstraction* (Kramsch, 1986: 367), and introduce far greater sampling complications for *general proficiency* than has been recorded for more specific language assessments (Shohamy, 1988:171-2; Chalhoub-Deville, 1997: 13). It seems that if we are to make holistic judgements to differentiate second language performance, more empirical guidance will be required than current descriptions (Buck, 1991:30-35; Fulcher, 1996: 208; Chalhoub-Deville, 1997: 9). An essential step towards this will be to further our understanding of *what actually happens* in live assessment encounters. This may be informed, along the way, by experimental and quantitative designs, but is most likely to develop significantly through collaborative data collections of qualitative researchers across a wide variety of contexts.

John Pollard, Riyadh, June 1999

References

Atkinson J M and Heritage J, 1984: *The interaction of talk with nonvocal activities*. In Atkinson and Heritage (Eds) 1984: 223-4.

Bachman L S, 1990: *Fundamental Principles in Language Testing*. OUP.

Bachman L S and Palmer A S, 1996: *Language Testing in Practice*. OUP.

Barnwell D, 1989: *'Naive' native speakers and judgements of oral proficiency in Spanish*. LT 6/2:152-63.

Brown A, 1998: *Interviewer style and candidate performance in the IELTS oral interview*. Paper given at the LTRC, Monterey, CA, March, 1998.

Buck G, 1991: *The Testing of Second Language Listening Comprehension*. Unpublished PhD Thesis, Lancaster University.

Chalhoub-Deville M, 1995: *A Contextualised Approach to Describing Oral Language Proficiency*. Language Learning 45/2:251-281.

Chalhoub-Deville M, 1997: *Theoretical models, assessment frameworks and test construction*. Language Testing 14/1:3-22.

Chapelle and Douglas, 1999: *Two new books on language test development*. Language Testing, 16/1:113-121.

Creswell J W, 1998: *Qualitative Inquiry and Research Design: Choosing Among Five Traditions*. Sage.

Cumming A, 1995: *Introduction: The Concept of Validation in Language Testing*. In Cumming and Berwick (eds): *Validation in Language Testing* 1-14. Multilingual Matters.

Eggins S and Slade D, 1997: *Analysing Casual Conversation*. Cassell.

Eisenstein M and Starbuck R, 1989: *The Effect of Emotional Investment on L2 Production*. In Gass, Madden, Preston and Selinker (eds), 1989: 125-140. Multilingual Matters.

Færch C and Kasper G. 1984. *Two ways of defining communication strategies.* Language Learning 34/1: 45-63.

Fulcher G. 1994. *Some priority areas for research in oral language testing.* Language Testing Update. 15:39-47.

Fulcher G. 1996. *Does thick description lead to smart tests ? A data-based approach to rating scale construction.* Language Testing. 13/2:208-38.

Goodwin C. 1984. *Notes on story structure and the organisation of participation.* In Atkinson and Heritage (Eds) 1984: 223-4.

Grice H P. 1975. *Logic and Conversation.* In Cole p and Morgan J L. eds. Syntax and Semantics 3: Speech Acts. Academic Press.

Heritage J. 1989. *Current developments in conversation analysis.* In Roger and Bull (eds) *Conversation.* Multilingual Matters. 21-47.

Hopper R. 1989. *Conversation analysis and social psychology as descriptions of interpersonal communication.* In Roger and Bull (eds). *Conversation.* Multilingual Matters. 48-65.

Jefferson G. 1989. *Preliminary notes on a possible metric which provides for a 'standard maximum' silence of approximately one second in conversation.* In Roger and Bull (eds) *Conversation.* Multilingual Matters. 166-196.

Johnson M and Tyler A. 1998. *Re-analyzing the OPI: How Much Does It look Like Natural Conversation ?* In Young and He (eds).1998.27-51. John Benjamins.

Jones E E and Gerard H B. 1967. *Foundations of Social Psychology.* John Wiley and Sons.

Johnson M and Tyler A. 1998. *Re-analyzing the OPI: How Much Does It Look like Natural Conversation ?*

Kramsch C. 1986. *From Language Proficiency to Interactional Competence.* The Modern Language Journal. 70/iv.

Kormos J. 1999. *Simulating conversations in oral-proficiency assessment: a conversation analysis of role plays and non-scripted interviews in language exams.* Language Testing16/2:163-188.

Lazaraton A. 1992. *The structural organisation of a language interview: a conversation analytic perspective.* System. 20: 373-86.

Lazaraton A. 1996a. *Interlocutor support in oral proficiency interviews : the case of CASE.* Language Testing 13/2: 151-172.

Lazaraton A. 1996b. *A qualitative approach to monitoring examiner conduct in the Cambridge assessment of spoken English (CASE).* In Milanovic M and Saville N (eds).1996. *Performance Testing, Cognition and assessment. Studies in Language Testing 3.* CUP.

Lazaraton A L. 1999. *Paper given at TESOL, New York on the proportions of qualitative versus quantitative studies published in four professional EFL Journals.*

Lynch K L and McNamara T F. 1998. *Using G-theory and Many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants.* Language Testing 15/2: 158-180.

Malabonga V. 1998. The Computerised Oral Proficiency Interview. LTU 24: 29, International Language Testers' Association.

McNamara T F. 1996. *Measuring Second Language Performance*. Longman.

McNamara T F and Lumley T. 1997. *The effect of interlocutor and assessment mode variables in overseas assessments of speaking skills in occupational settings*. Language Testing, 14/2: 140-156.

Messick S. 1994. *The Interplay of Evidence and Consequences in the Validation of Performance Assessments*. Educational Researcher 23/2:13-23.

Milanovic M, Saville N, Pollitt A and Cook A.1995. *Developing Rating Scales for CASE: Theoretical Concerns and Analyses*. In Cumming and Berwick (eds): Validation in Language Testing. 15-38. Multilingual Matters.

Munby J. 1978. *Communicative Syllabus Design*. CUP.

Pollard J D E and Underhill N. 1996. *Developing and Researching Validity for a Computer-Resourced Proficiency Interview Test*. Language Testing Update 20. 49-52. ILTA.

Pollard J D E. 1998. *Research and development – a complex relationship*. Language Testing Update 24. 46-57. ILTA.

Porter D. 1991. *Affective factors in the assessment of oral interaction: gender and status*. In Anivan S. 1991 (ed) Current developments in language testing. Singapore: SEAMEO RELC. 99-102.

Ross S. 1992. *Accommodative questions in oral proficiency interviews*. Language Testing 9/2:173-186.

Ross S and Berwick R. 1992. *The discourse of accommodation in oral proficiency examinations*. STUDIES IN SECOND LANGUAGE ACQUISITION (SSLA). 14:159-76.

Shegloff E A. 1995. *Discourse as an Interactional Achievement III: The Omnirelevance of Action*. In Jacoby and Ochs (eds): *Special issue on co-construction*. Research on Language and Social Interaction 28/3. 171-83.

Shohamy E. 1988. *A Proposed Framework for Testing the Oral Language of Second/Foreign Language Learners*. STUDIES IN SECOND LANGUAGE ACQUISITION (SSLA). 10:165-179. CUP.

Shohamy E. 1994. *The validity of direct versus semi-direct oral tests*. Language Testing. 11/ 2. 1994 99-123.

Silverman D. 1993. *Interpreting Qualitative Data: Methods for Analysing Talk, Text and Interaction*. Sage.

Skehan P. 1990. *Progress in Language Testing in the 1990s*. in Alderson and North (eds) *Language Testing in the 1990s*. Modern English Publications and British Council.

Strauss A and Corbin J. 1990. *Basics of Qualitative Research: Grounded Theory Procedures and Techniques*. Sage.

Tannen D. 1989. *Talking Voices: Reptition, Dialogue and Imagery in Conversational Discourse*. CUP.

Tarone E. 1980. *Communication Strategies, Foreigner Talk, and Repair in Interlanguage*. Language Learning 30. 417-431.

BEST COPY AVAILABLE

Tarone E. 1982: *Systematicity and attention in interlanguage*. Language Learning 32: 69-82.

Tarone E. 1983: *Some thoughts on the notion of 'communication strategy'*. In Faerch and Kasper (eds) 1983:61-74. Longman.

Underhill N. 1987: *Testing Spoken Language*. Cambridge University Press.

Underhill N. 1997: *Unpublished research report*.

Upshur J and Turner C E. 1999: *Systematic effects in the rating of second-language speaking ability: test method and learner discourse*. Language Testing 16/1 82-111.

Van Lier L. 1989: *Reeling, Writhing, Drawling, Stretching, and Fainting in Coils: Oral Proficiency Interviews as Conversation*. TESOL Quarterly. 23/3.

White R. 1997: *Back channelling, repair, pausing, and private speech*. Applied Linguistics 18/3: 314-344.

Wigglesworth G. 1993: *Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction*. Language Testing 10/3: 305-335.

Wilkins D. 1976: *Notional Syllabuses*. OUP.

Woken M D and Swales J. 1989: *Expertise and authority in native-non-native conversations: The need for a variable account*. In Gass, Madden, Preston and Selinker (eds): *Variation in Second Language Acquisition: Discourse and Pragmatics*. Multilingual Matters.

Young R. 1995: *Conversational Styles in language Proficiency Interviews*. Language Learning 45/1: 3-42.

Young R and Milanovic M. 1992: *Discourse variation in oral proficiency interviews*. Studies in Second Language Acquisition (SSLA): 14:403-24.

Young R and He A W. 1998: *Language Proficiency Interviews: A Discourse Approach. Introduction: Assessing Second Language Speaking Ability*. In Young and He (eds): Talking and Testing. John Benjamin: 1-24.

Zeungler J. 1989: Performance variation in NS-NNS interactions: Ethnolinguistic difference, or discourse domain ? In Gass, Madden, Preston and Selinker (eds): *Variation in Second Language Acquisition: Discourse and Pragmatics*. Multilingual Matters.